Agenda

- Should we trust the results?
- What are the results telling us about education in the state?
- How can we use the results to improve education in the state?
- What resources are we providing to educators and students to help target instruction?

Should we trust the results?

- Background—Has the test or passing score changed?
    - McCrea's claim
    - Evidence
- What do the early years of a testing program typically look like?
    - Stability and change—sources of variation in any test
    - Typical patterns and comparisons with other Smarter Balanced and non-Smarter Balanced states
- Summary

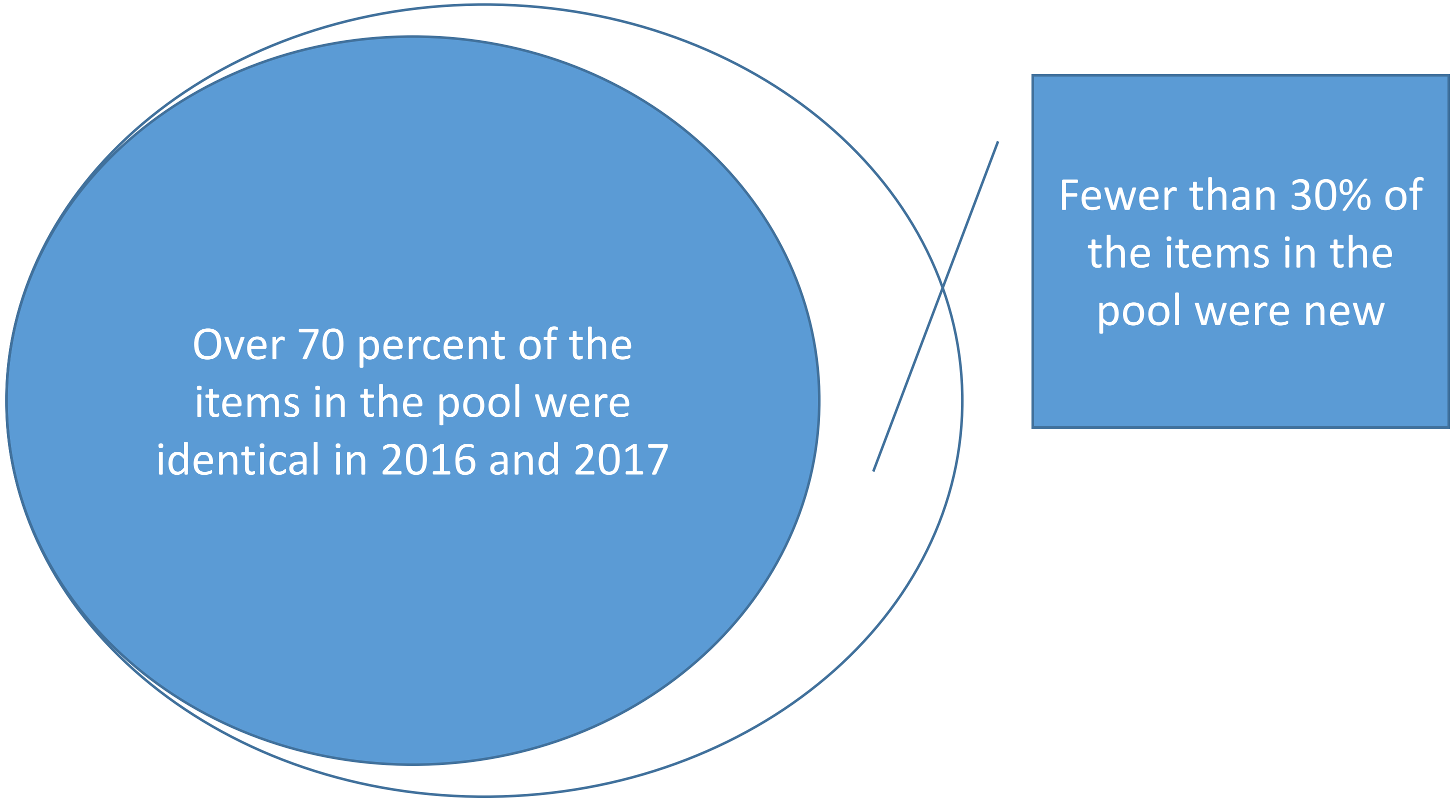
McRea’s Claim: Smarter Balanced states declined, while PARCC states improved or stayed the same

Table 1: General pattern of change over years, Smarter Balanced and PARCC

Subject	Year	Smarter Balanced	PARCC
ELA	2015-16	⬆	➡
	2016-17	⬇	⬆
Math	2015-16	⬆	⬆
	2016-17	➡	➡

1. *McRea calls it “Fair Game” to assign letter grades based on no-change constituting failure (F), and extraordinarily high gains (4 points) an A. This choice makes the pattern in Table 1 seem more extreme.*
2. *McRea casts suspicions on the expansion of the Smarter Balanced item pool.*

Did the newly introduced items introduce a downward bias?

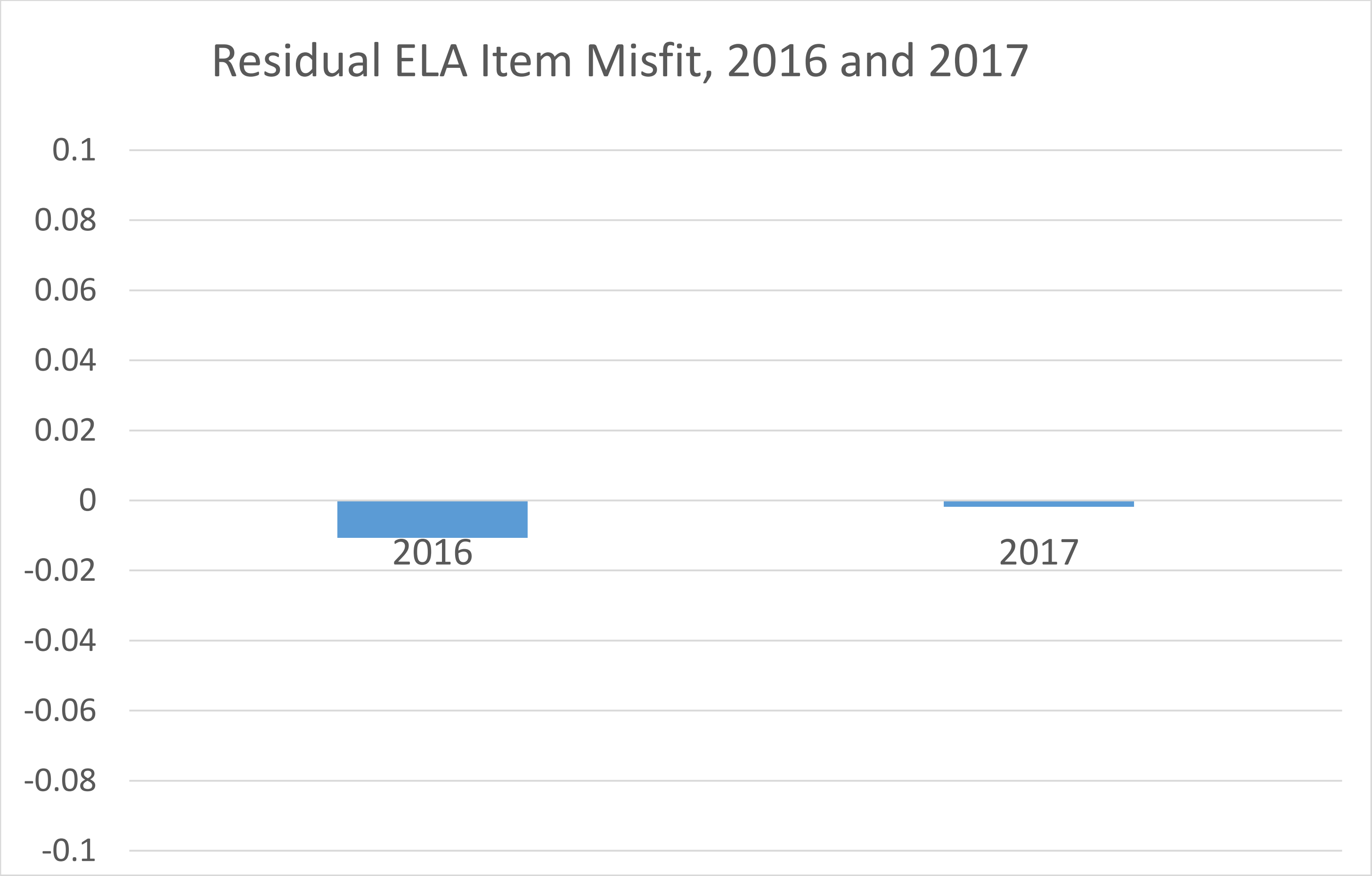


Over 70 percent of the items in the pool were identical in 2016 and 2017

Fewer than 30% of the items in the pool were new

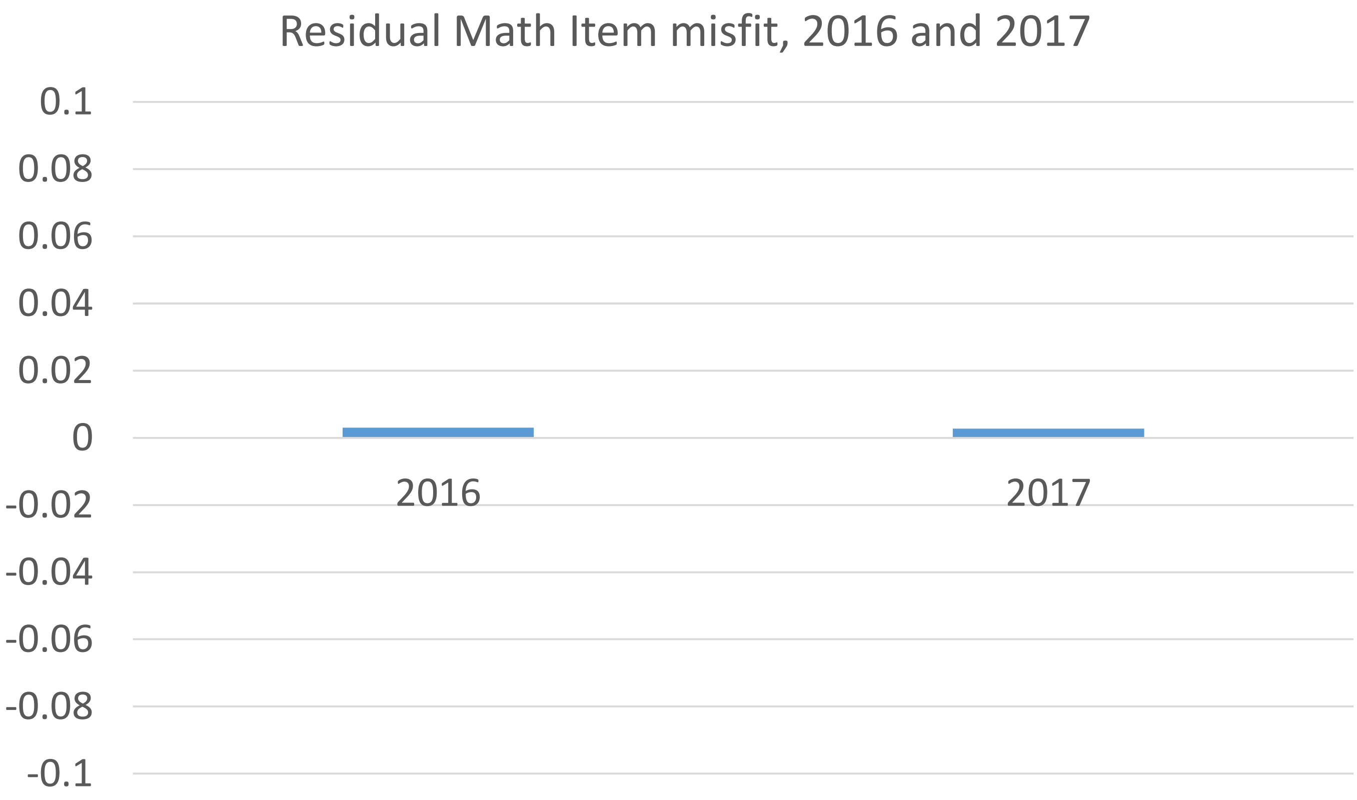
Unlikely, since 70% of the items in the pool were unchanged from 2016-2017.

But did the new items function differently?



No. The items functioned almost exactly as expected. The items that were ***common*** *across years proved trivially more difficult than expected. The* ***new*** *items functioned as expected, and were not a source of bias.*

Same story in math: Items performed as expected, new and old



Sources of change in statewide test scores over time

- Changing cohorts of students. In Vermont, you would expect a minimum of 0.5-1.0 change in the percent proficient just due to sampling error.
  - Even this assumes that stability in terms of demographics, student experience, etc.
- Variation due to the items on a test
  - Equating variance can be large on a fixed form test, where a small number of items is used to link this year’s test to last year’s
  - Equating variance is much, much smaller on adaptive tests, which typically maintain most of a much larger pool from year to year
  - A study in Ohio a few years ago found that some linking procedures can lead to substantial shifts of several percentage points in the percent proficient.
- True changes in student performance

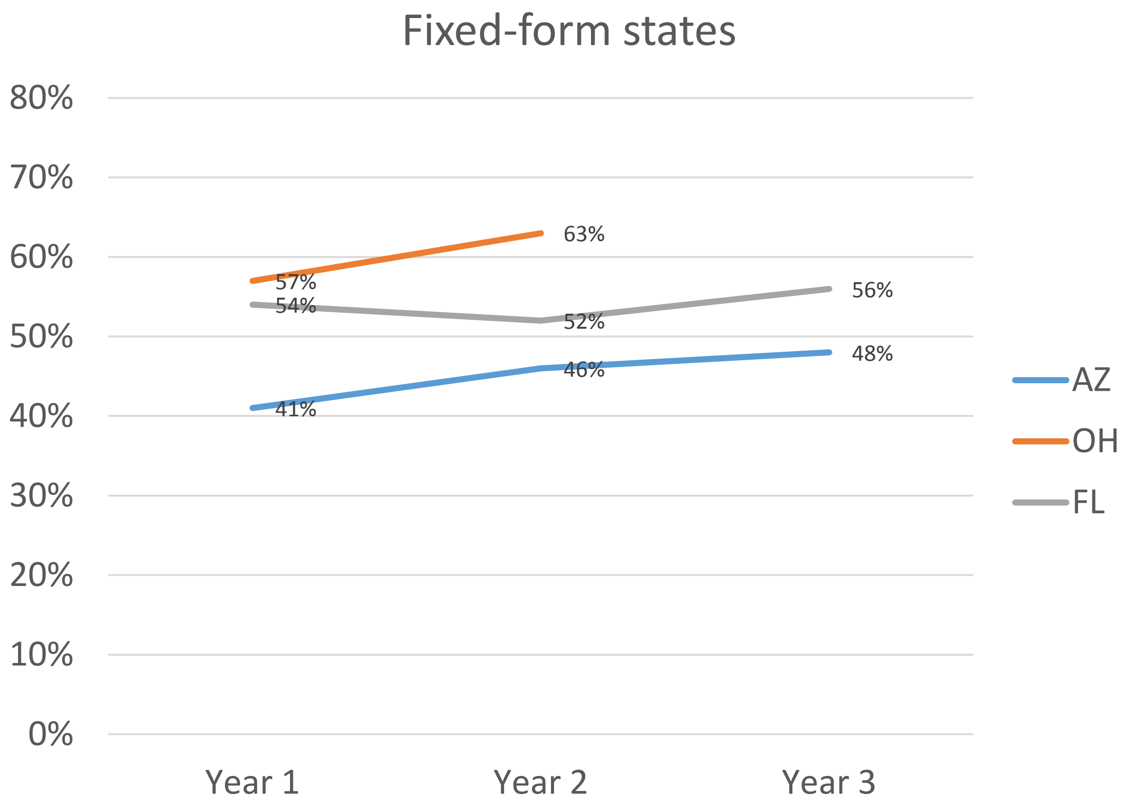
So what do the early years of a testing program look like?

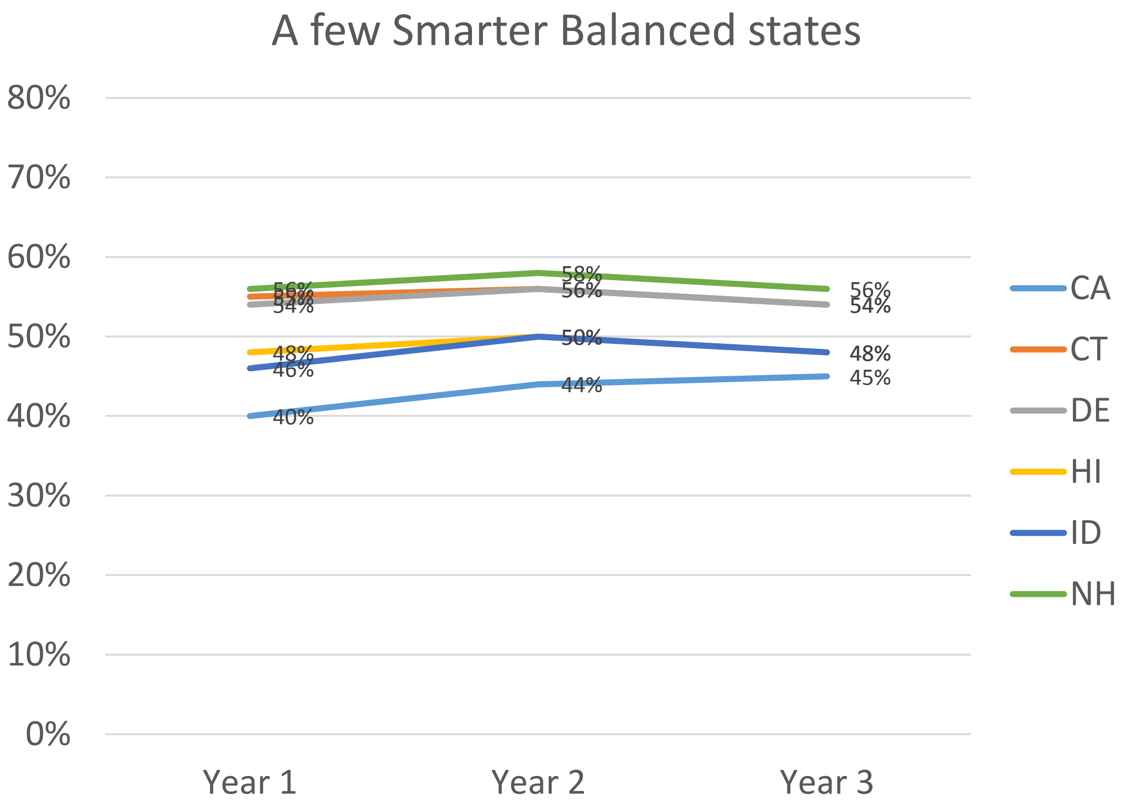
- Comparing three groups of AIR clients that started new testing programs in 2014-15 or 2015-16
  - **Fixed-form states**: Arizona, Ohio, and Florida
  - **Six Smarter Balanced states** for comparison (limited to keep the graphs readable)
  - **Vermont and Utah**, because Utah started an independent adaptive testing program and therefore makes a good comparison for Vermont.

What patterns will we see in the data?

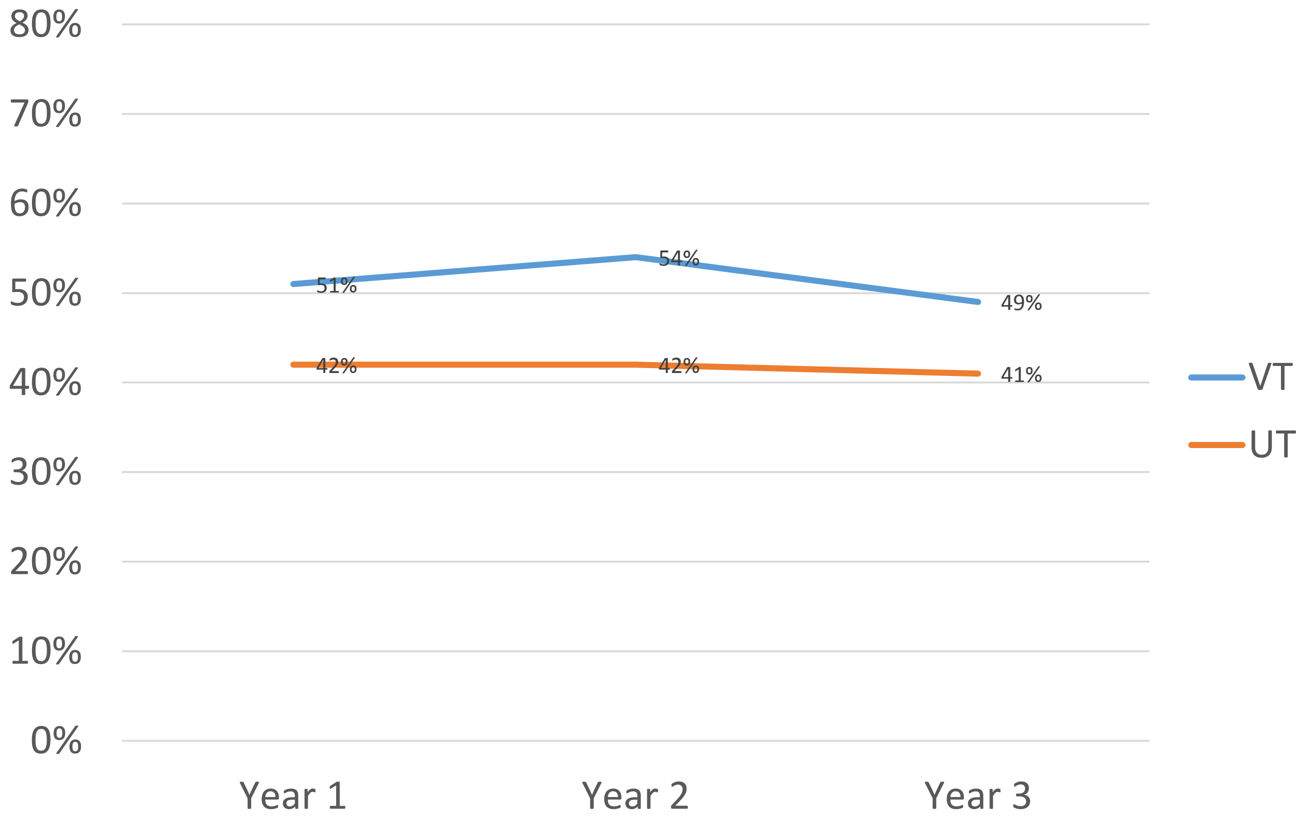
- Typically, growth the first year, followed by leveling off subsequent years
- Fixed-form tests show *larger changes* than adaptive tests
    - They are subject to substantially more linking error, so there is simply more noise in the year-to-year data
    - Our example includes larger states, so the volatility due to sampling of students across cohorts is lower
    - A greater proportion of the variance is likely due to equating variance than in Vermont or the Smarter Balanced states

Percent proficient over time from program inception, Grade 4 ELA

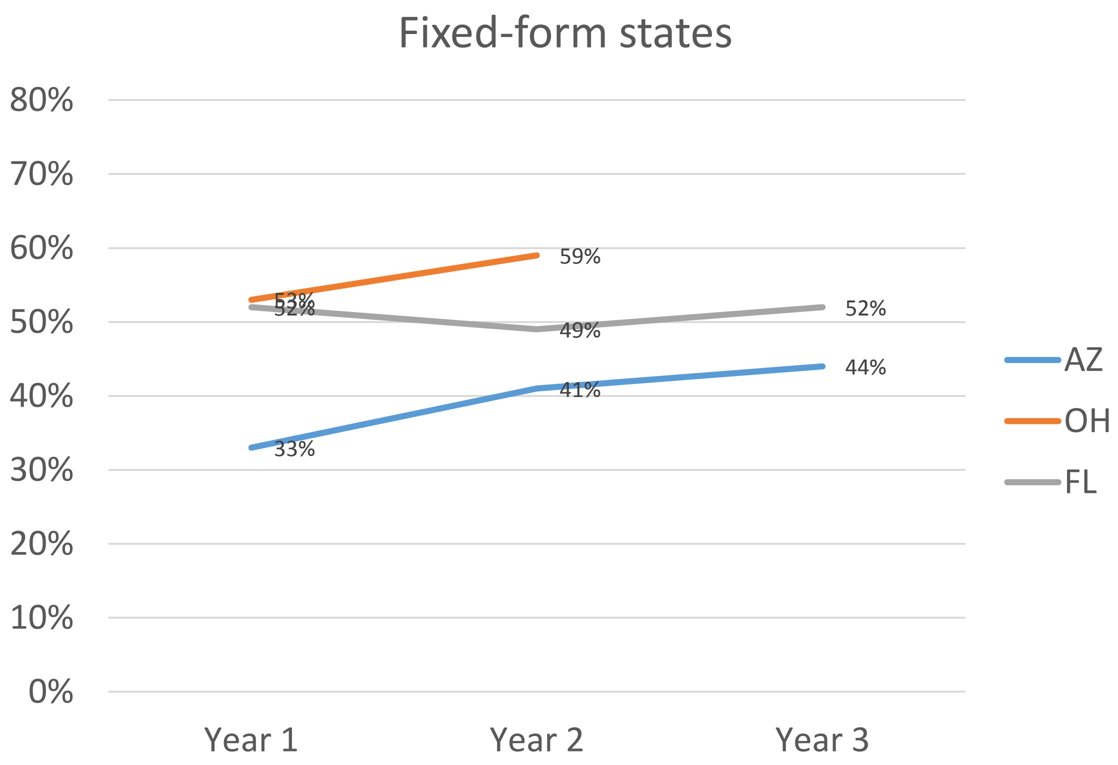


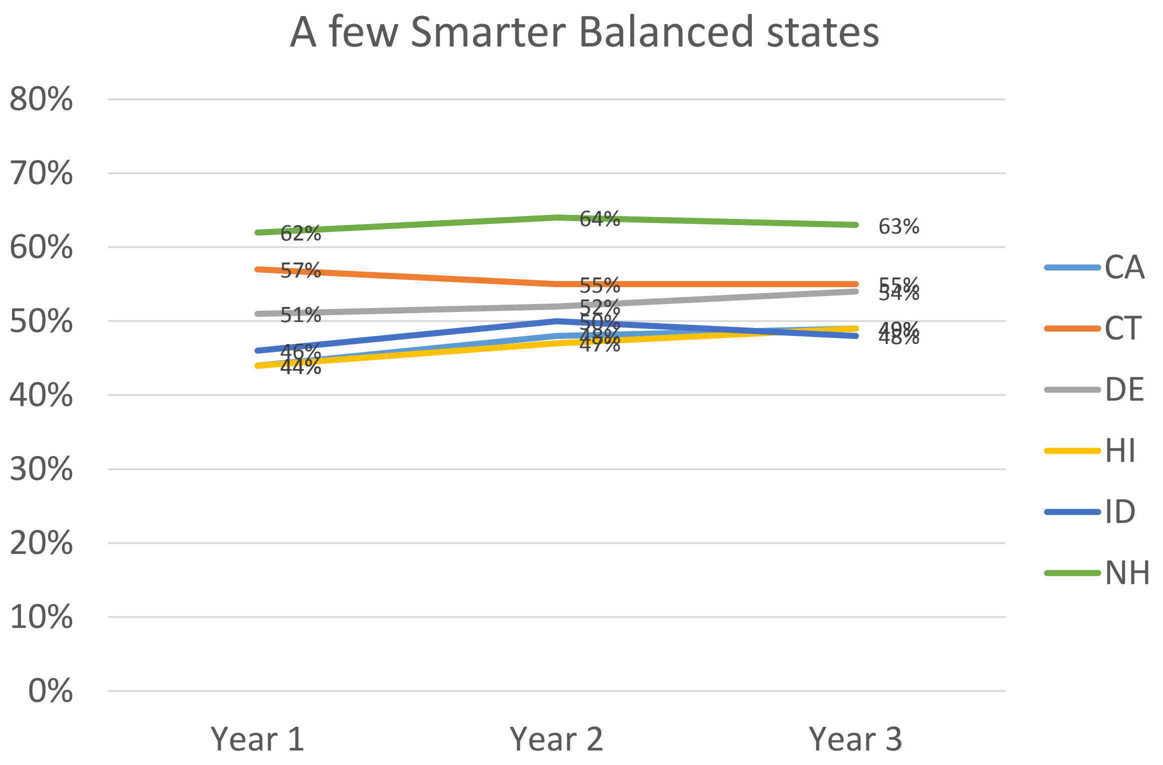


Percent proficient over time from program inception, Grade 4 ELA: Utah and Vermont

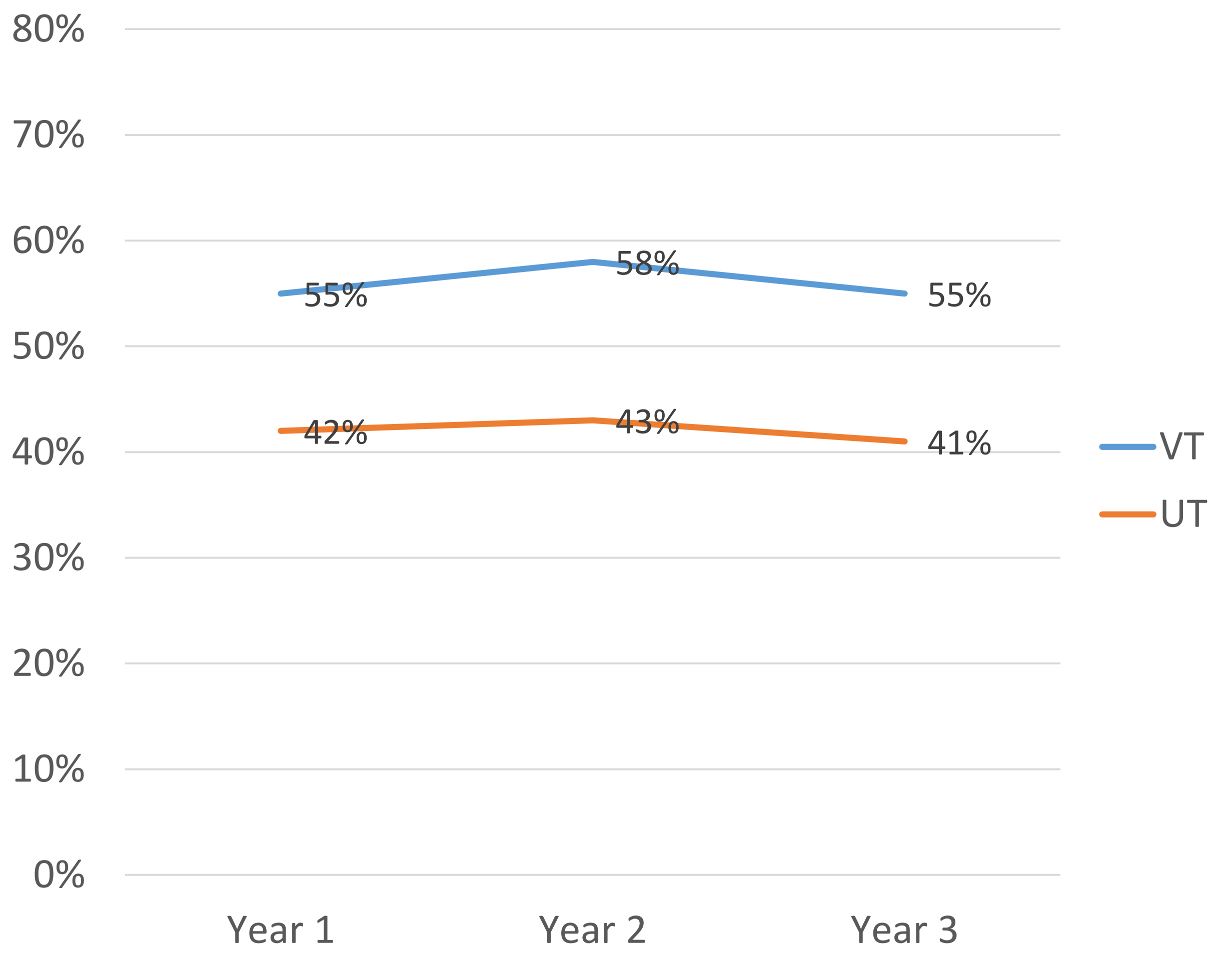


Percent proficient over time from program inception, Grade 7 ELA

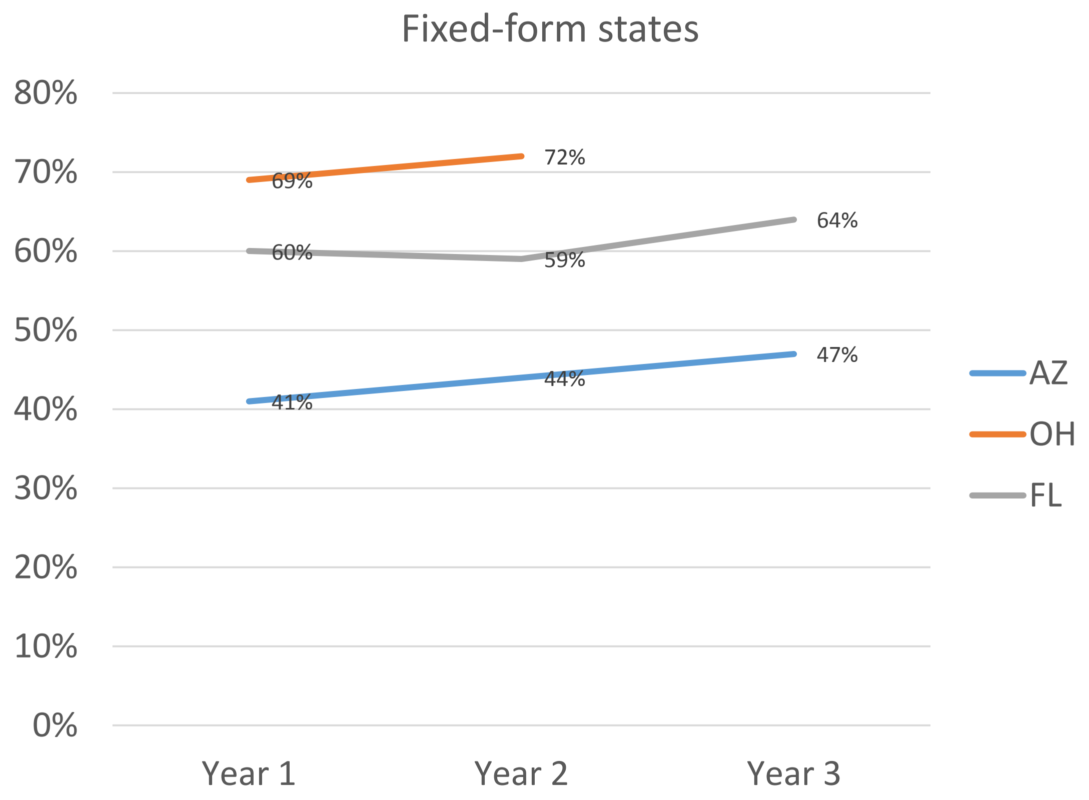

Fixed-form states

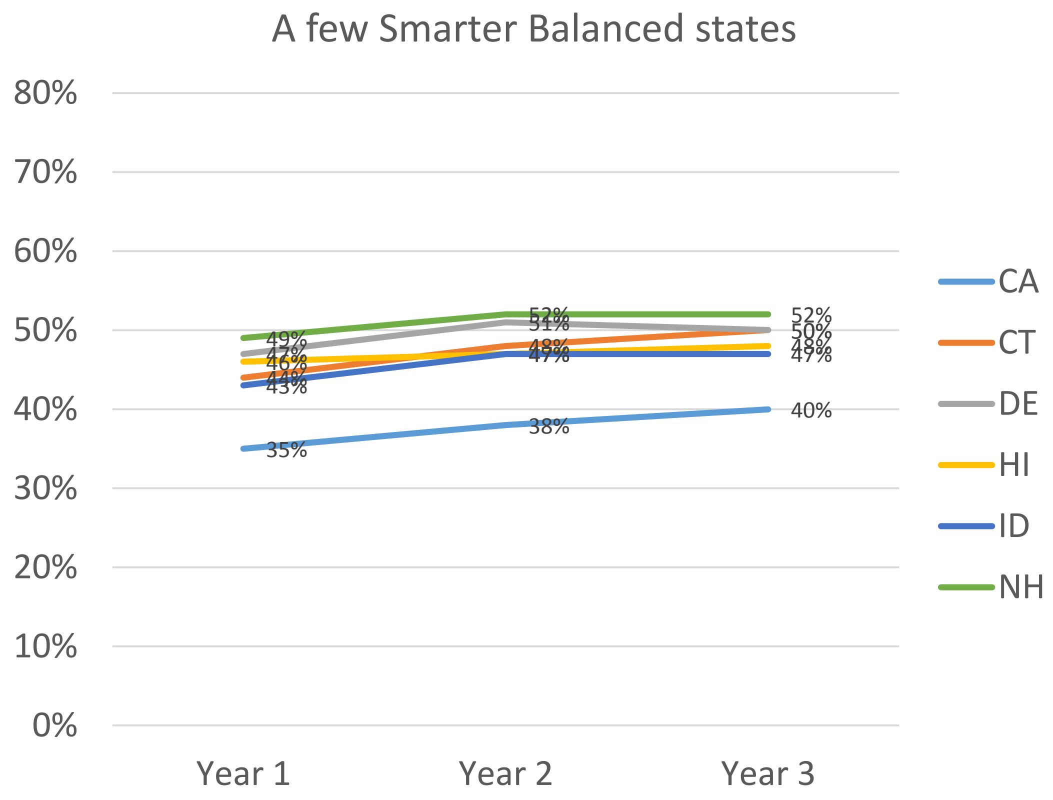

A few Smarter Balanced states

Percent proficient over time from program inception, Grade 7 ELA: Utah and Vermont

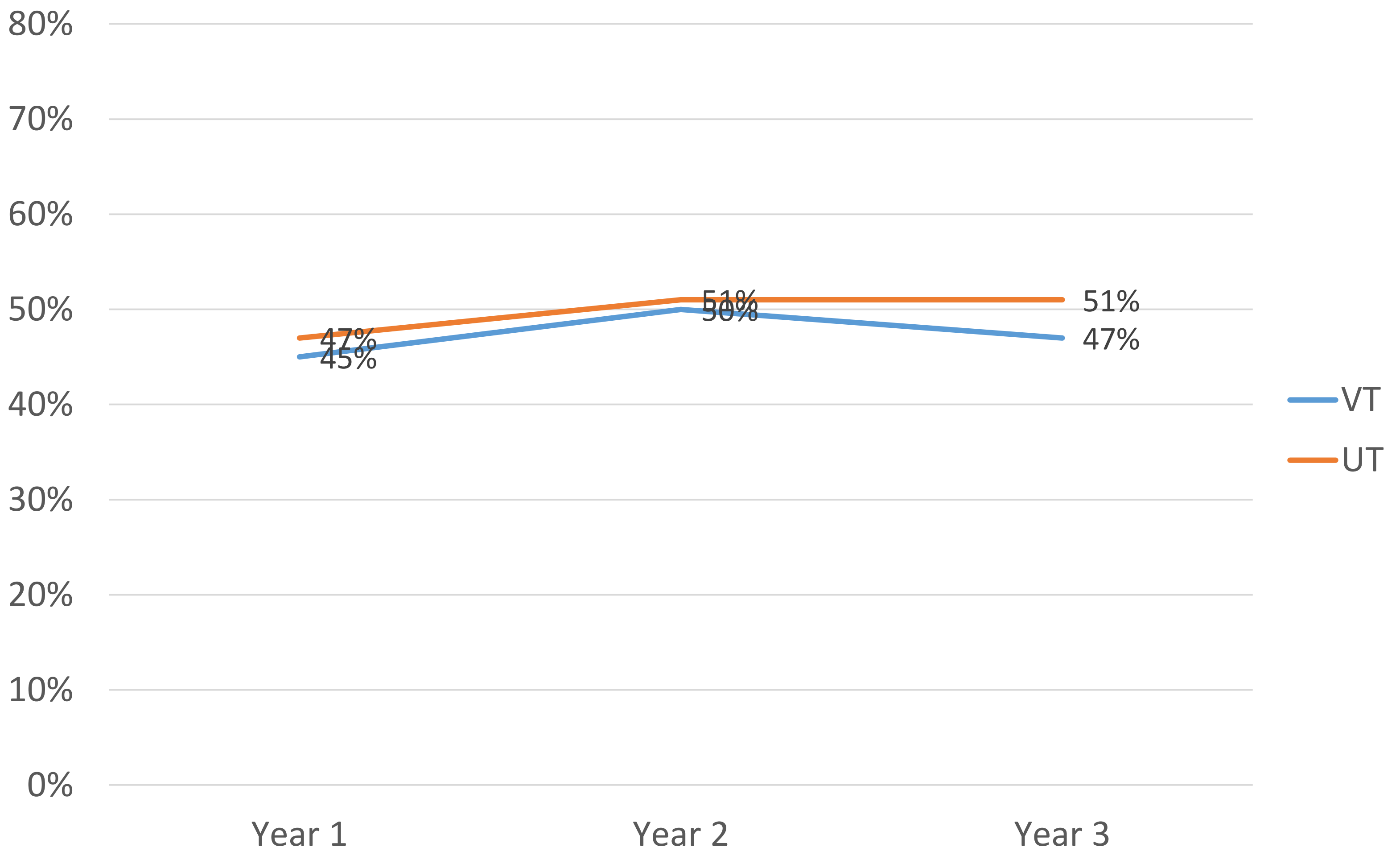


Percent proficient over time from program inception, Grade 4 Math

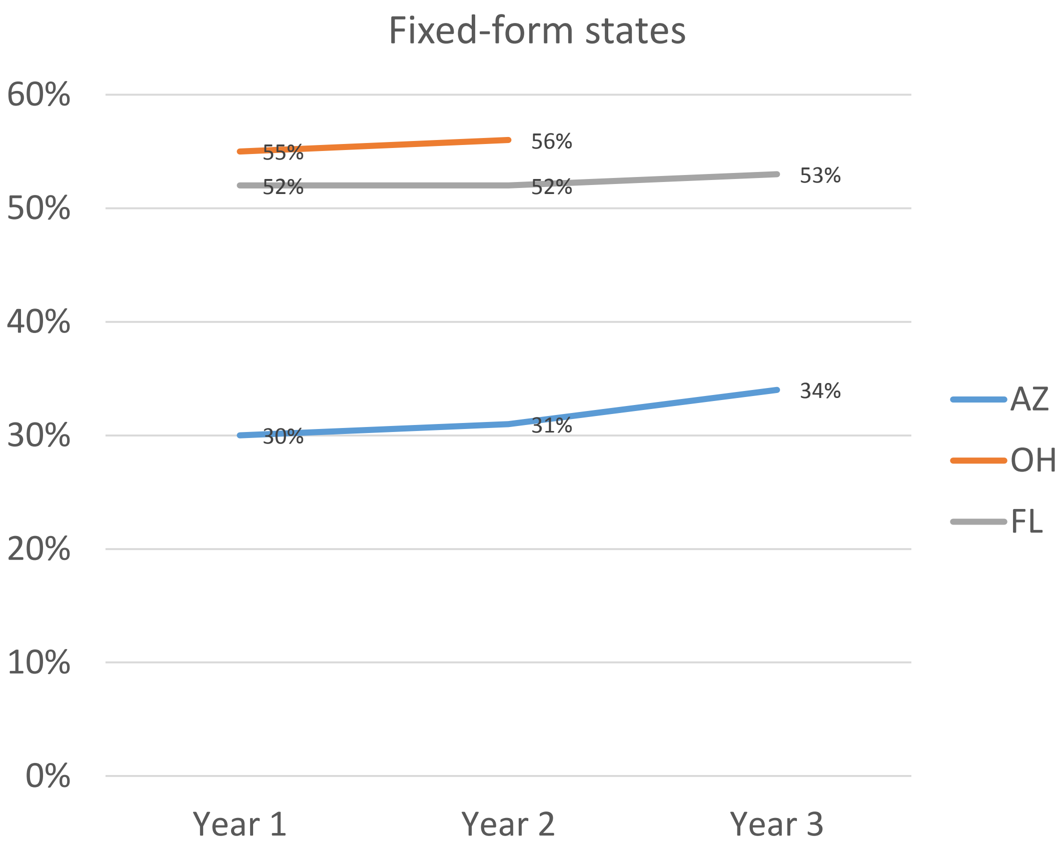


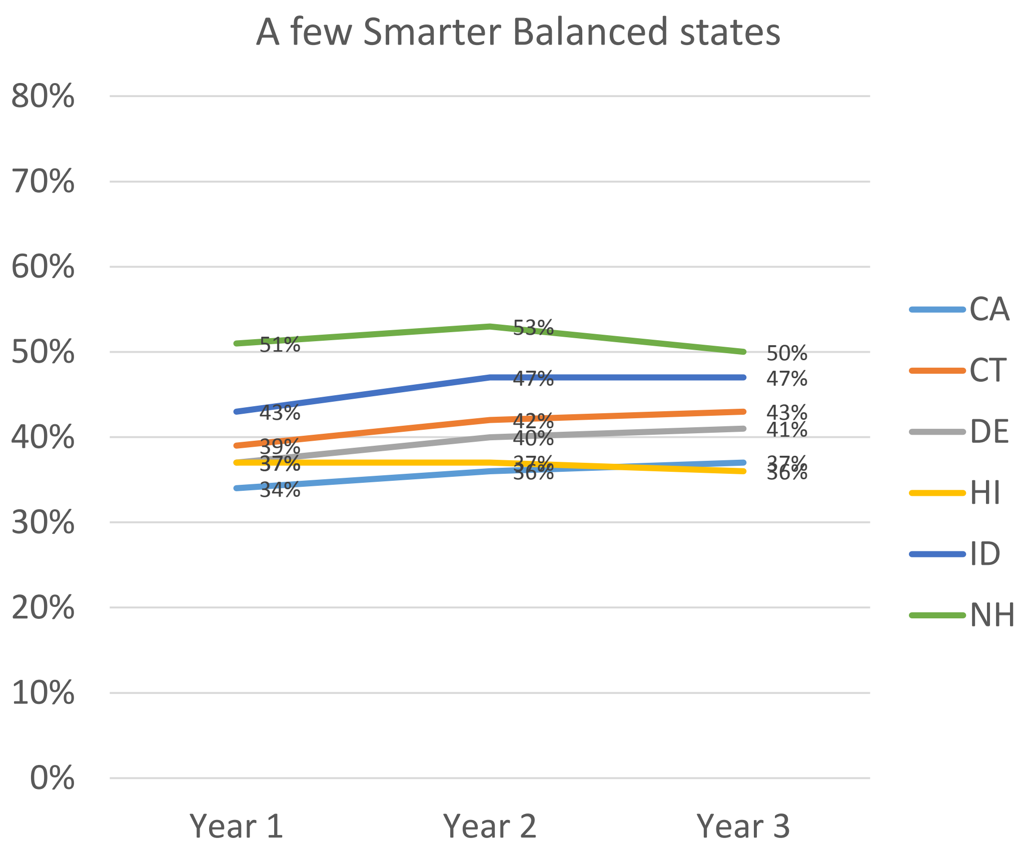


Percent proficient over time from program inception, Grade 4 Math: Utah and Vermont

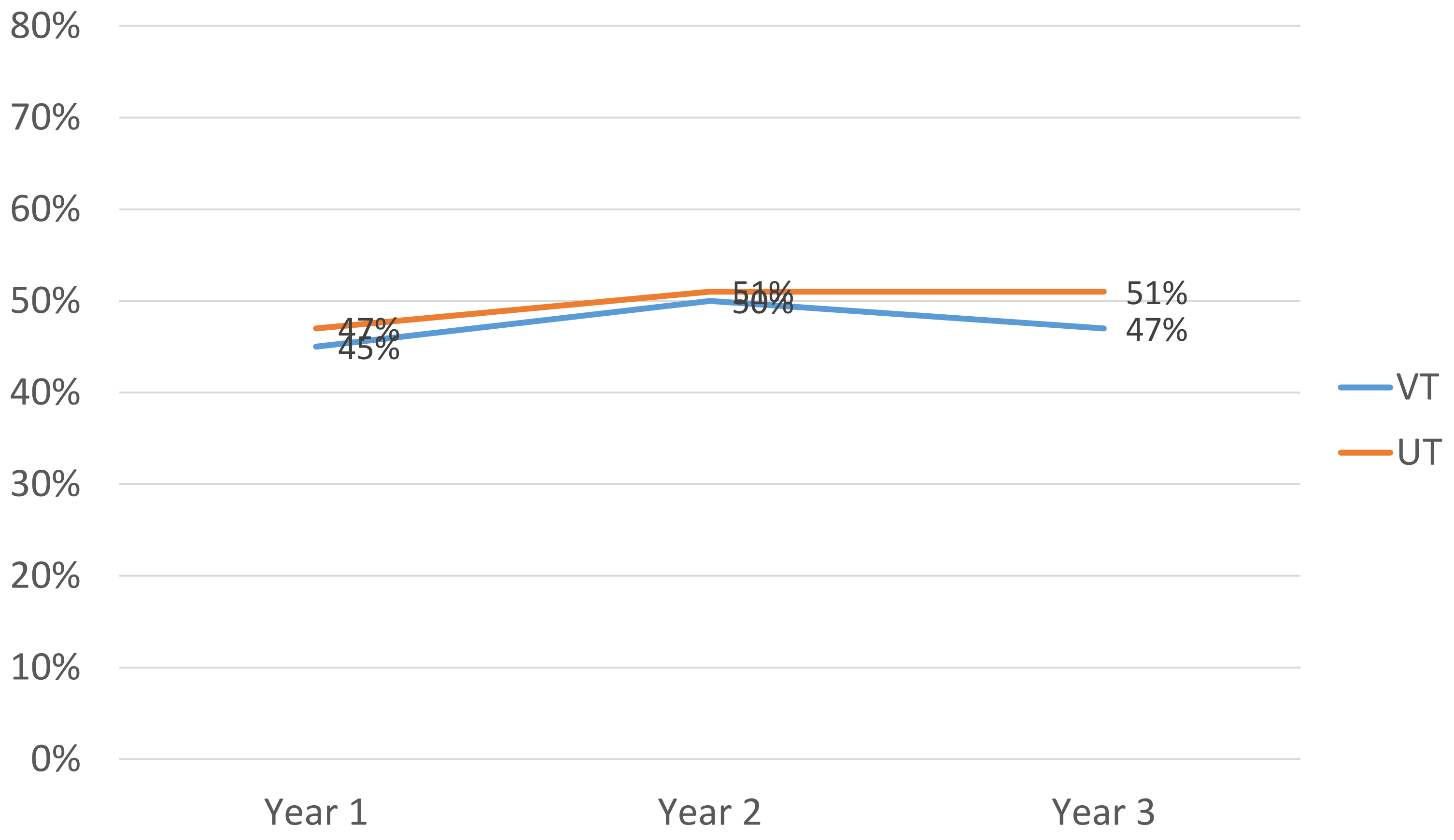


Percent proficient over time from program inception, Grade 7 Math


Fixed-form states


A few Smarter Balanced states

Percent proficient over time from program inception, Grade 7 Math: Utah and Vermont



Summary

- Expect somewhat bigger random shifts from fixed-form states than from Smarter Balanced and other adaptive states due to equating variance
- Typical pattern shows substantial increase from Year 1 to Year 2, with a subsequent leveling off
- The data is behaving as expected, in the absence of substantial changes in student learning.

What are the results telling us?

What do the results tell us

- Vermont has shown very small improvements from 2015-2017
- There is little evidence of substantial educational change in the state over that time.
  - Typical boost between 2015 and 2016.
  - Leveling off or slight decline in 2016-2017.

How can we use the test results to improve education

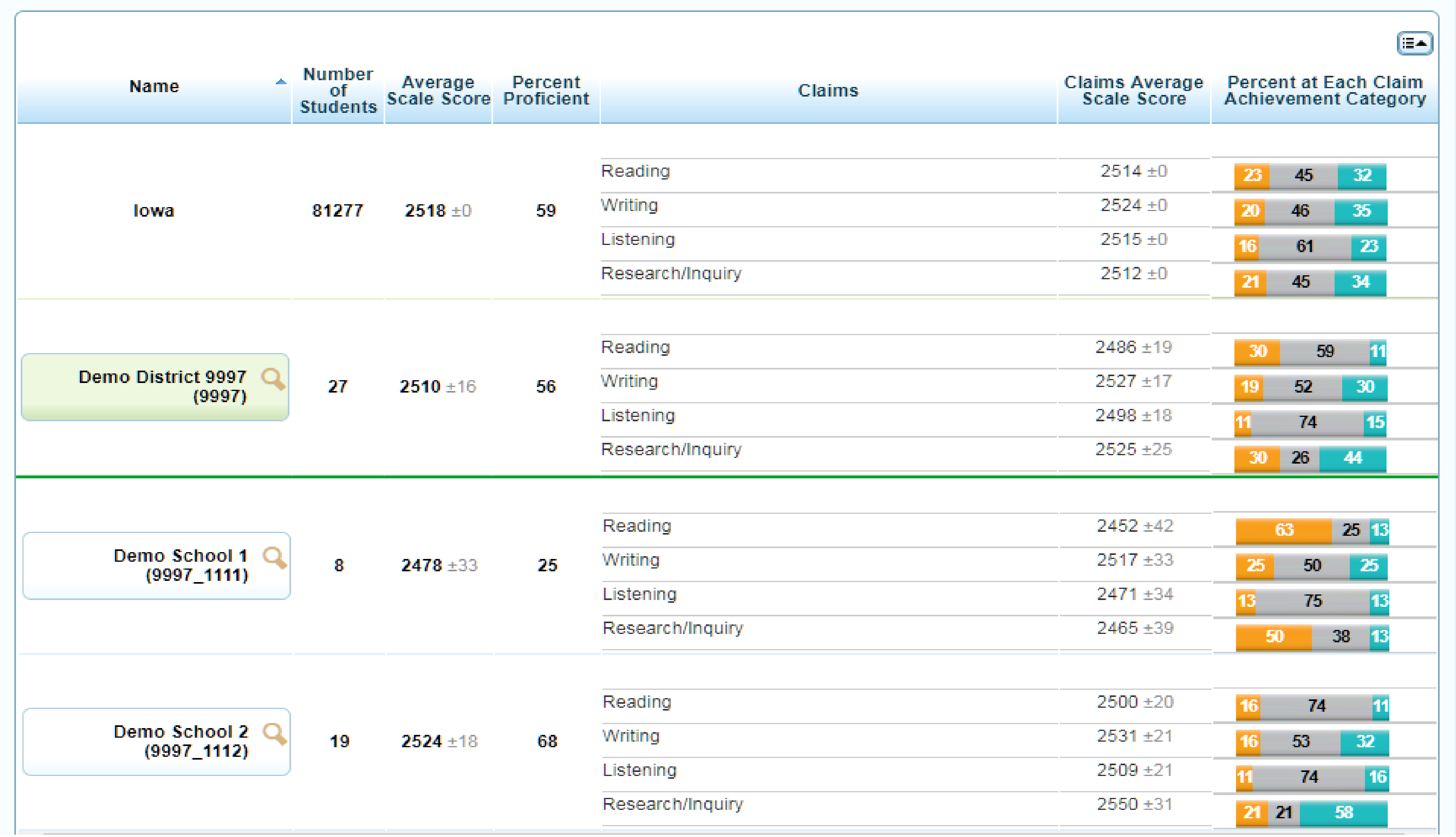
State-level uses

- Audits and Accountability
  - Multi-tiered system of supports is currently self-reported. Where reported implementation does not correspond with improved test scores, maybe dig in deeper.
  - One measure in an accountability system that includes some consequences.
- Program evaluation-keep what works and improve what does not.
  - Evaluate whether student’s rate of learning increases among students of teachers who take advantage of professional learning opportunities
    - Help identify those that are not effective
    - Help steer educators towards those that are
  - Evaluate contracts with school turnaround and other consultants

District, school, and teacher uses

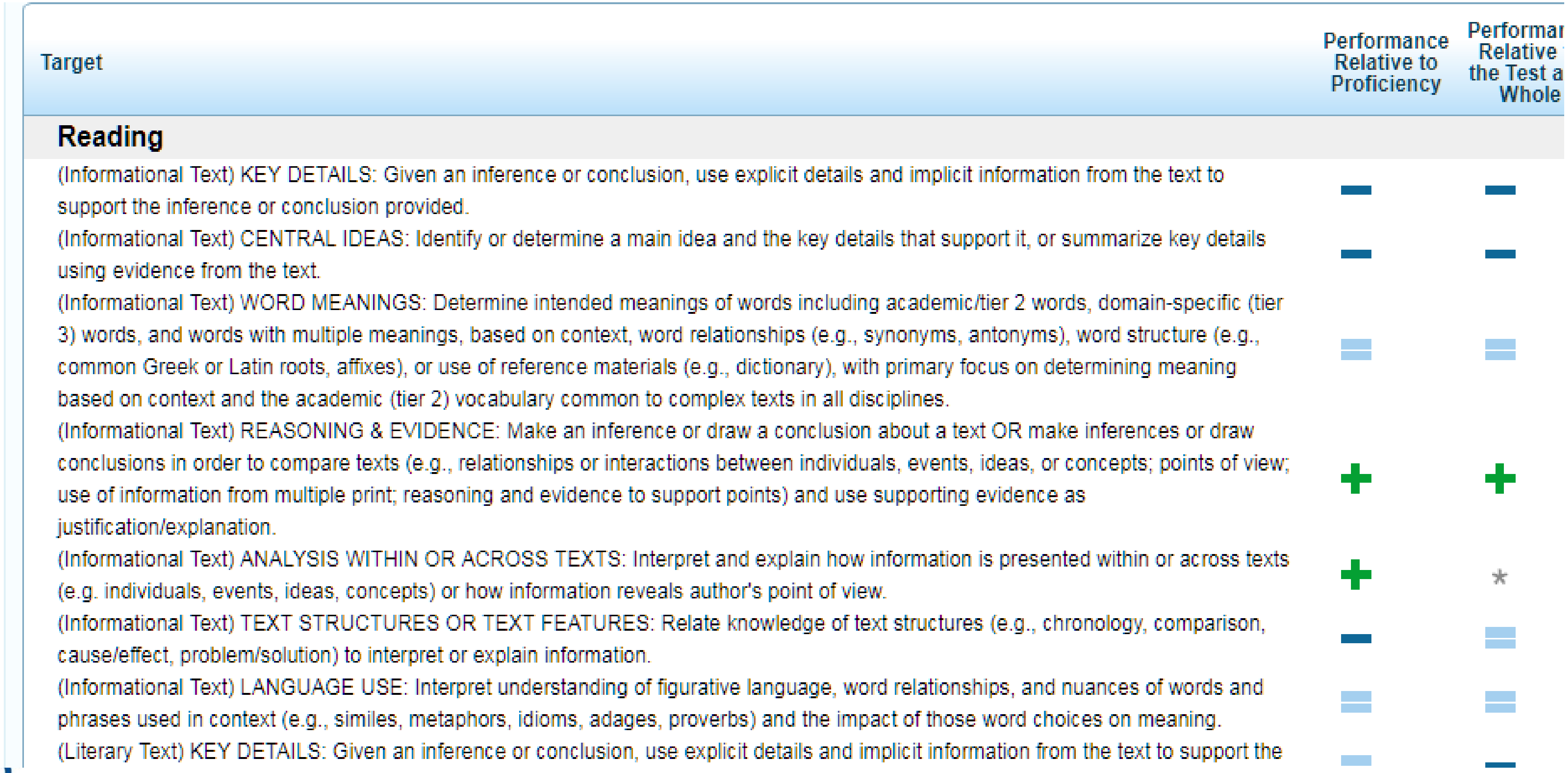
- Interactive reporting system enables educators to
  - Track customized groups of students, including classes, subgroups within or across classes
- Identify what is working in the curriculum or classroom

Detailed reporting by Claim, district, school, classroom, other grouping



Name	Number of Students	Average Scale Score	Percent Proficient	Claims	Claims Average Scale Score	Percent at Each Claim Achievement Category
Iowa	81277	2518 ±0	59	Reading	2514 ±0	23 45 32
				Writing	2524 ±0	20 46 35
				Listening	2515 ±0	16 61 23
				Research/Inquiry	2512 ±0	21 45 34
Demo District 9997 (9997)	27	2510 ±16	56	Reading	2486 ±19	30 59 11
				Writing	2527 ±17	19 52 30
				Listening	2498 ±18	11 74 15
				Research/Inquiry	2525 ±25	30 26 44
Demo School 1 (9997_1111)	8	2478 ±33	25	Reading	2452 ±42	63 25 13
				Writing	2517 ±33	25 50 25
				Listening	2471 ±34	13 75 13
				Research/Inquiry	2465 ±39	50 38 13
Demo School 2 (9997_1112)	19	2524 ±18	68	Reading	2500 ±20	16 74 11
				Writing	2531 ±21	16 53 32
				Listening	2509 ±21	11 74 16
				Research/Inquiry	2550 ±31	21 21 58

Detailed reporting by Target, district, school, classroom, other grouping



Target	Performance Relative to Proficiency	Performar Relative the Test a Whole
Reading		
(Informational Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.	−	−
(Informational Text) CENTRAL IDEAS: Identify or determine a main idea and the key details that support it, or summarize key details using evidence from the text.	−	−
(Informational Text) WORD MEANINGS: Determine intended meanings of words including academic/tier 2 words, domain-specific (tier 3) words, and words with multiple meanings, based on context, word relationships (e.g., synonyms, antonyms), word structure (e.g., common Greek or Latin roots, affixes), or use of reference materials (e.g., dictionary), with primary focus on determining meaning based on context and the academic (tier 2) vocabulary common to complex texts in all disciplines.	=	=
(Informational Text) REASONING & EVIDENCE: Make an inference or draw a conclusion about a text OR make inferences or draw conclusions in order to compare texts (e.g., relationships or interactions between individuals, events, ideas, or concepts; points of view; use of information from multiple print; reasoning and evidence to support points) and use supporting evidence as justification/explanation.	+	+
(Informational Text) ANALYSIS WITHIN OR ACROSS TEXTS: Interpret and explain how information is presented within or across texts (e.g. individuals, events, ideas, concepts) or how information reveals author's point of view.	+	*
(Informational Text) TEXT STRUCTURES OR TEXT FEATURES: Relate knowledge of text structures (e.g., chronology, comparison, cause/effect, problem/solution) to interpret or explain information.	−	=
(Informational Text) LANGUAGE USE: Interpret understanding of figurative language, word relationships, and nuances of words and phrases used in context (e.g., similes, metaphors, idioms, adages, proverbs) and the impact of those word choices on meaning.	=	=
(Literary Text) KEY DETAILS: Given an inference or conclusion, use explicit details and implicit information from the text to support the	=	−

Summary

Question	Answer
Can we trust the results or are there issues with calibration or linking?	The test results are stable, valid, and reliable, and accurately reflect learning.
What pattern of improvement do we expect when a new test is introduced?	What we see in Vermont is pretty typical.
What are the results telling us?	We are not seeing the improvement that we would like to see.
What can the state do?	Use the testing data for a strong accountability system, to target audits for your educational improvement programs, to evaluate the efficacy of programs such as professional development offerings and other educational improvement initiatives. Keep what works, and replace what does not.
What can educators do?	Use the reported results to evaluate curricula, teaching methods, etc. to see what works and replace things that do not. Use the data to identify groups of students with specific skills or deficits to target instruction more effectively.